

RESEARCH ARTICLE

Long-time analytic approximation of large stochastic oscillators: Simulation, analysis and inference

Giorgos Minas^{1,2}, David A. Rand^{1,2*}

1 Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, University of Warwick, Coventry, United Kingdom, **2** Mathematics Institute, University of Warwick, Coventry, United Kingdom

* d.a.rand@warwick.ac.uk



Abstract

In order to analyse large complex stochastic dynamical models such as those studied in systems biology there is currently a great need for both analytical tools and also algorithms for accurate and fast simulation and estimation. We present a new stochastic approximation of biological oscillators that addresses these needs. Our method, called phase-corrected LNA (pcLNA) overcomes the main limitations of the standard Linear Noise Approximation (LNA) to remain uniformly accurate for long times, still maintaining the speed and analytical tractability of the LNA. As part of this, we develop analytical expressions for key probability distributions and associated quantities, such as the Fisher Information Matrix and Kullback-Leibler divergence and we introduce a new approach to system-global sensitivity analysis. We also present algorithms for statistical inference and for long-term simulation of oscillating systems that are shown to be as accurate but much faster than leaping algorithms and algorithms for integration of diffusion equations. Stochastic versions of published models of the circadian clock and NF- κ B system are used to illustrate our results.

OPEN ACCESS

Citation: Minas G, Rand DA (2017) Long-time analytic approximation of large stochastic oscillators: Simulation, analysis and inference. PLoS Comput Biol 13(7): e1005676. <https://doi.org/10.1371/journal.pcbi.1005676>

Editor: Bard Ermentrout, University of Pittsburgh, UNITED STATES

Received: November 2, 2016

Accepted: July 6, 2017

Published: July 24, 2017

Copyright: © 2017 Minas, Rand. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was funded by the BBSRC Grant BB/K003097/1 (Systems Biology Analysis of Biological Timers and Inflammation). DAR was also supported by funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 305564. BBSRC web site: www.bbsrc.ac.uk Seventh Framework Programme (FP7) website: cordis.europa.eu/fp7/home_en.html. The funders had no role in study

Author summary

Many cellular and molecular systems such as the circadian clock and the cell cycle are oscillators that are modelled using nonlinear dynamical systems. Moreover, oscillatory systems are ubiquitous elsewhere in science. There is an extensive theory for perfectly noise-free dynamical systems and very effective algorithms for simulating their temporal behaviour. On the other hand, biological systems are inherently stochastic and the presence of stochastic noise can play a crucial role. Unfortunately, there are far fewer analytical tools and much less understanding for stochastic models especially when they are nonlinear and have lots of state variables and parameters. Moreover simulation is not so effective and can be very slow if the system is large. In this article we describe how to accurately approximate such systems in a way that facilitates fast simulation, parameter estimation and new approaches to analysis, such as calculating probability distributions that describe the system's stochastic behaviour and describing how these distributions change when the parameters of the system are varied.

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Dynamic cellular oscillating systems such as the cell cycle, circadian clock and other signaling and regulatory systems have complex structures, highly nonlinear dynamics and are subject to both intrinsic and extrinsic stochasticity. Moreover, current models of these systems have high-dimensional phase spaces and many parameters. Modelling and analysing them is therefore a challenge, particularly if one wants to both take account of stochasticity and develop an analytical approach enabling quantification of various aspects of the system in a more controlled way than is possible by simulation alone. The stochastic kinetics that arise due to random births, deaths and interactions of individual species give rise to Markov jump processes that, in principle, can be analyzed by means of master equations. However, these are rarely tractable and although an exact numerical simulation algorithm is available [1], for the large systems we are interested in, this is very slow.

It is therefore important to develop accurate approximation methods that enable a more analytical approach as well as offering faster simulation and better algorithms for data fitting and parameter estimation. A number of approximation methods aimed at accelerating simulation are currently available. This includes leaping algorithms [2, 3] and algorithms for integration of diffusion equations (or chemical Langevin equations (CLE)) [4] that provide faster simulation. However, these methods do not provide analytical tools for studying the dynamics of the system and they can also be extremely slow for data fitting and parameter estimation. One obvious candidate for overcoming these limitations is the Linear Noise Approximation (LNA). The LNA is based on a systematic approximation of the master equation by means of van Kampen's Ω -expansion [5] which uses the system size parameter Ω that controls the number of molecules present in the system. The large system size validity of the LNA has been shown in [6], in the sense that the distribution of the Markov jump process at a fixed finite time converges, as the system size Ω tends to ∞ , to the LNA probability distribution. The latter distribution is analytically tractable allowing for fast estimation and simulation algorithms. However, the LNA has significant limitations, particularly in approximating long-term behaviour of oscillatory systems.

We show below that for the oscillatory systems that we study, the LNA approximation of the distribution $P_t = P(Y(t)|Y(0))$, of the state $Y(t)$ of the system at some time t becomes inaccurate when the time t is greater than a few periods of the oscillation. However, if we rather consider a similar system which in the $\Omega \rightarrow \infty$ limit instead of a limit cycle has an equilibrium point that is linearly stable, then the LNA approximation of P_t remains accurate for a much longer time-scale. For example, in Fig C in [S3 Appendix](#) we give an example where the LNA fails in a matter of a period or two for the oscillatory system, but for the corresponding equilibrium system it is very accurate for over a hundred times as long (and probably much longer). Similar behaviour is also observed in other systems and using different measures in [8].

The observation that non-degenerate limit cycles have such linearised stability in the directions transversal to the limit cycle suggests the way forward for oscillatory systems. Our approach exploits the fact that, because of this transversal linearised stability, the distributions P_t for a general class of systems with a stable attracting limit cycle in the $\Omega \rightarrow \infty$ limit are, like the above fixed point systems, similarly well-behaved on long time-scales provided one conditions P_t on appropriate transversal sections to the limit cycle.

We introduce a modified LNA, called the phase-corrected LNA, or pcLNA, that exploits the above observations to overcome the most important shortcomings of the LNA and we develop methods for analysis, simulation and inference of oscillatory systems that are accurate for much larger times. We build on previous work of Boland et al. [9] which uses the 2-dimensional Brusselator system as an exemplar to investigate the failure of the LNA in

approximating long-term behaviour of oscillatory systems and presents a method for computing power spectra and comparing exact simulations with LNA predictions of the same phase rather than time. Using various low-dimensional oscillatory systems for illustration, a related analysis has been employed to study the temporal variability of oscillatory systems in the tangential direction of the $\Omega \rightarrow \infty$ limit cycle [10] and/or the amplitude variability in the transversal direction of the limit cycle [11–13]. Other papers derive related descriptions of the asymptotic phase of stochastic oscillators [14, 15].

We extend these results in a number of ways including the following: (i) we develop a theory that treats the general case and provide analytical arguments that justify our approximations and enable computation of trajectory distributions, (ii) we show that the approach is practicable for larger nonlinear systems, (iii) we present a new powerful system-global sensitivity theory for such systems using measures such as the Fisher Information Matrix and the Kullback-Leibler divergence that are analytically computed, (iv) we present a simulation algorithm and show it is as accurate but faster than leaping and integration of diffusion equation algorithms, and (v) we derive the Kalman filter associated with the pcLNA in order to provide a practical way to accurately approximate the likelihood function thus facilitating estimation of system parameters θ and predictive algorithms. The approach in [9] uses transversal sections which are normal to the limit cycle. We follow this but in the supplementary information (S1 Sects. 8.2 & 8.3) we show that for most considerations one can use any transversal to the limit cycle, including those defined in [14, 15].

To illustrate and validate our approach we apply it to a relatively large published stochastic model of the *Drosophila* circadian clock due to Gonze et al. [16] (see S2 Sect. 1). This model involves 10 state variables and 30 reactions and its structure is discussed in S2. The large system limit is given by the differential equation system of 10 kinetic equations that are listed in the supplementary information (S2 Sect. 1) along with the reaction scheme of the system. The stochastic version of the Brusselator system and a stochastic version of a well-studied model of the NF- κ B signalling system [17] are also used to illustrate our methods and the results can be found in S3 Appendix.

These systems are free-running oscillators in the sense that they correspond to a limit cycle of an autonomous differential equation in the $\Omega \rightarrow \infty$ limit. However, our results also apply to the equally important classes of entrained forced oscillators and damped oscillations. We therefore consider two such systems in S2 and S3: the light-entrained *Drosophila* circadian clock model of [18] which is an example of a forced oscillator and the NF- κ B system model [17]. The latter has the extra feature that the analysis is not concerned with a limit cycle but of a transient solution that converges to the limit cycle as time increases. This solution is the biologically interesting one that describes how the system responds to being stimulated by TNF α . The supplementary information S1 includes technical derivations and S2 and S3 contain further illustrative figures that we refer to in this paper.

Results

Stochastic models of cellular processes in signaling and regulatory systems are usually described in terms of reaction networks. A system of multiple different molecular subpopulations has state vector, $Y(t) = (Y_1(t), \dots, Y_n(t))^T$ where $Y_i(t)$, $i = 1, \dots, n$, denotes the number of molecules of each species at time t . These molecules undergo a number of possible reactions (e.g. transcription, translation, degradation) where the reaction of index j changes $Y(t)$ to $Y(t) + v_j$, $v_j \in \mathbb{R}^n$. The vectors v_j are called stoichiometric vectors. Each reaction occurs randomly at a rate $w_j(Y(t))$ (often called the intensity of the reaction), which is a function of $Y(t)$

but also depends periodically on t when we are studying forced oscillators. Such systems can be exactly simulated using the Stochastic Simulation algorithm (SSA) [1].

It is common in studying stochastic systems to introduce a system size Ω which is a parameter that occurs in the intensities of the reactions $w_j(Y(t))$ and controls molecular numbers (see discussion in S1 Sect. 2). While having a system size parameter is not necessary to apply our methods, it allows one to study the dependence of stochastic fluctuations upon system size and to calculate the deterministic equations that describe the evolution of the concentration vector $X(t) = Y(t)/\Omega$ in the limit of $\Omega \rightarrow \infty$ (see S1 Sect. 3). Although more general conditions can be used, a condition that will be sufficient for our purposes is that the rates $w_j(Y(t))$ depend upon Ω as $w_j(Y) = \Omega u_j(Y/\Omega)$ (cf. [5–7]).

In the limit $\Omega \rightarrow \infty$ the time dependence of $X(t)$ is given by the solution $x(t)$ of the differential equation

$$\dot{x} = F(x), \quad F(x) = \sum_j v_j u_j(x(t)), \quad (1)$$

with the appropriate initial condition (see S1 Sect. 3). For free-running oscillators the differential equation Eq (1) is autonomous, whereas for forced oscillators F also depends periodically on t .

Throughout we will be interested in the case where the solution $x(t)$ of interest is a stable limit cycle γ of minimal period $\tau > 0$ given by $x = g(t)$, $0 \leq t \leq \tau$. We shall also always assume the generic situation for stable limit cycles of autonomous systems in which one of the characteristic exponents of the limit cycle is equal to zero and the rest have negative real part ([19], [20 Sect 1.5] and S1 Sect. 1). For an entrained forced oscillator all the characteristic exponents of the limit cycle are assumed to have negative real parts.

Fig 1(A) displays a stochastic trajectory of the concentrations $X(t) = Y(t)/\Omega$ of two of the species of the circadian clock system obtained using exact SSA simulation over a period of time $t \in [0, 8.5\tau]$ where $\tau \approx 26.98$ hours is the period of the limit cycle γ . Here the system size is $\Omega = 300$, imposing moderate to high levels of stochasticity (see also Table C in S2 Appendix). Results for system sizes $\Omega = 200, 500$ and 1000 are also reported in Fig B in S2 Appendix. Fig 1 (B) shows realizations of the key probability distributions

$$P(X(t)|X(t_0) = x_0), \quad t > t_0, \quad (2)$$

which describe the state of the system at some time $t > t_0$. It is very rare that accurate analytical approximations for such probability distributions can be derived from the exact Markov-jump process when t is large. Furthermore, as we can see, the SSA samples of $P(X(t)|X(t_0) = x_0)$ spread along the curved periodic solution, $x = g(t)$, of the limiting ($\Omega \rightarrow \infty$) deterministic system, implying that for large t this distribution is far from being normal and has a complex structure.

However, we will show that there are important related distributions that can be well approximated even for large times t . For example, for each point x on the $\Omega \rightarrow \infty$ limit cycle γ , consider the $(n-1)$ -dimensional hyperplane \mathcal{S}_x normal to γ at $x \in \gamma$ (i.e. orthogonal to the tangent vector $F(x)$ at x). We will show that the distribution of the intersection points of stochastic trajectories $X(t)$ with \mathcal{S}_x can be well-approximated by a multivariate normal (MVN) distribution that can be relatively easily calculated.

We need to define more precisely what we mean by intersection points. Consider the mapping G_N defined by $G_N(X) = x \in \gamma$ if $X \in \mathcal{S}_x$ (see Fig 1(C)). Suppose that the limit cycle is given by $x = g(t)$ and extend $g(t)$ to all $t \in \mathbb{R}$ by periodicity. Now consider a stochastic trajectory $X(t_i)$, $i = 0, 1, 2, \dots$. Suppose that $G_N(X(t_i - 1)) = g(s_{i-1})$, $i > 0$. If $G_N(X(t_i)) = g(s)$ then it equals $g(s + q\tau)$ for all integers q . Choose q so that $s_i = s + q\tau$ satisfies $s_{i-1} - \tau/2 \leq s_i < s_{i-1} + \tau/2$.

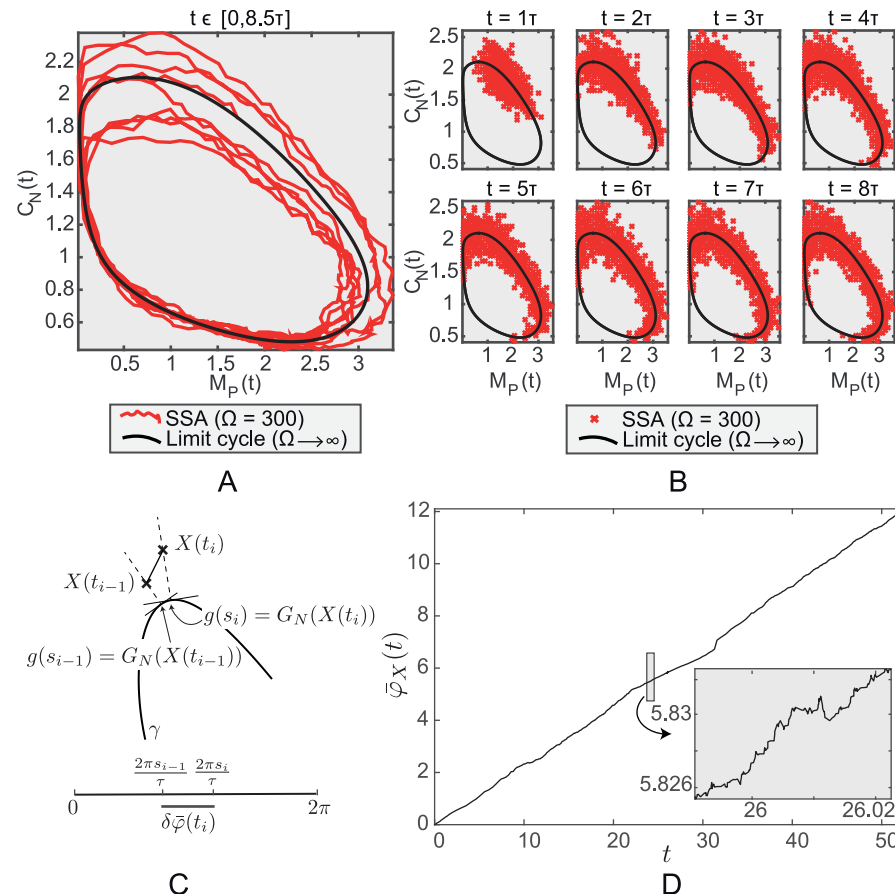


Fig 1. Exact stochastic simulation of the *Drosophila* circadian clock system. (A) A stochastic trajectory obtained by running the SSA over the time-interval $t \in [0, 8.5\tau]$ and (B) SSA samples ($R = 3000$) at times $t = \tau, 2\tau, \dots, 8\tau$. The concentrations, $X(t) = Y(t)/\Omega$, of two (out of 10) of the species are displayed (per mRNA M_P (x-axis) and nuclear PER-TIM complex C_N (y-axis)). For this model the system size Ω is Avogadro's number in units of nM^{-1} multiplied by the cell volume in litres and the concentrations are nanomolar. The value of Ω used here is 300 in units of nM which assumes a cell volume of approximately 0.5×10^{-12} litres. The number of M_P molecules ranges from 0 to 1200 and the number of C_N molecules from 100 to 900. It is only through Ω that the cell volume appears in the equations. The black solid curve is the $\Omega \rightarrow \infty$, limit cycle solution. (C) Schematic diagram illustrating the mapping $G_N(\cdot)$ and the relative phase $\delta\varphi_X(\cdot)$. (D) A plot of the lifted phase function $\bar{\varphi}_X(t)$ for a trajectory of this system. Note the long-term linear increase combined with frequent reversals.

<https://doi.org/10.1371/journal.pcbi.1005676.g001>

Note that the difference between s_i and s_{i-1} can be both positive and negative and the choice of s_i minimises $|s_i - s_{i-1}|$. We define the relative phase $\delta\varphi_X(t_i)$ to be $2\pi(s_i - s_{i-1})/\tau$. With this definition the lifted phase $\bar{\varphi}_X$ is defined to be the function

$$\bar{\varphi}_X(t_0) = 0 \quad \text{and} \quad \bar{\varphi}_X(t_i) = \sum_{k=1}^i \delta\varphi_X(t_k) = 2\pi(s_i - s_0)/\tau \quad \text{if} \quad i \geq 1.$$

An example of $\bar{\varphi}_X(t)$ is shown in Fig 1(C). Now, as shown in that figure, if the system size is not too small, although there will be some reversals, the stochastic process $G_N(X(t))$, $t > 0$, will move around γ in the direction of the deterministic flow given by Eq (1) so that $\bar{\varphi}_X(t)$ increases at a definite positive rate because phase advances exceed retreats. Our approximations (S1 Sect. 6) give that the long term rate is $2\pi/\tau$ and that the variance of the fluctuations $\bar{\varphi}_X(t) - 2\pi t/\tau$ grows linearly with a growth rate that is $O(t/\Omega)$.

Now consider stochastic trajectories $X(t_i)$ with $X(t_0) \in \mathcal{S}_{g(t_0)}$ and consider how this trajectory passes through $\mathcal{S}_{g(t_1)}$. We can assume that $T_0 < T_1 \leq T_0 + \tau$ by the periodicity of g . The r th pass will occur between the first time t_i when

$$\bar{\varphi}_X(t_i) \leq 2\pi(r-1) + 2\pi(T_1 - T_0)/\tau < \bar{\varphi}_X(t_{i+1})$$

and t_{i+1} . Therefore, we define the *first intersection point of the r th pass* to be

$$Q_{x_1}^{(r)} = Q_{x_1}^{(r,X)} = X(t_i)$$

since X does not change between t_i and t_{i+1} .

These points of intersection $Q_x^{(r)}$ describe the stochasticity of the system around a particular phase x of the system. The above ideas work equally well with any transversal to γ . For example, the time at which the i -th variable is maximal in the deterministic system is given by the transversal submanifold Σ defined by $\dot{x}_i = 0$, $\ddot{x}_i < 0$ so the intersection of the stochastic trajectory with Σ can be regarded as giving the statistics of the maxima of the stochastic trajectory (see Fig 2(A)). Close to the limit cycle, Σ is well-approximated by its tangent space at the point

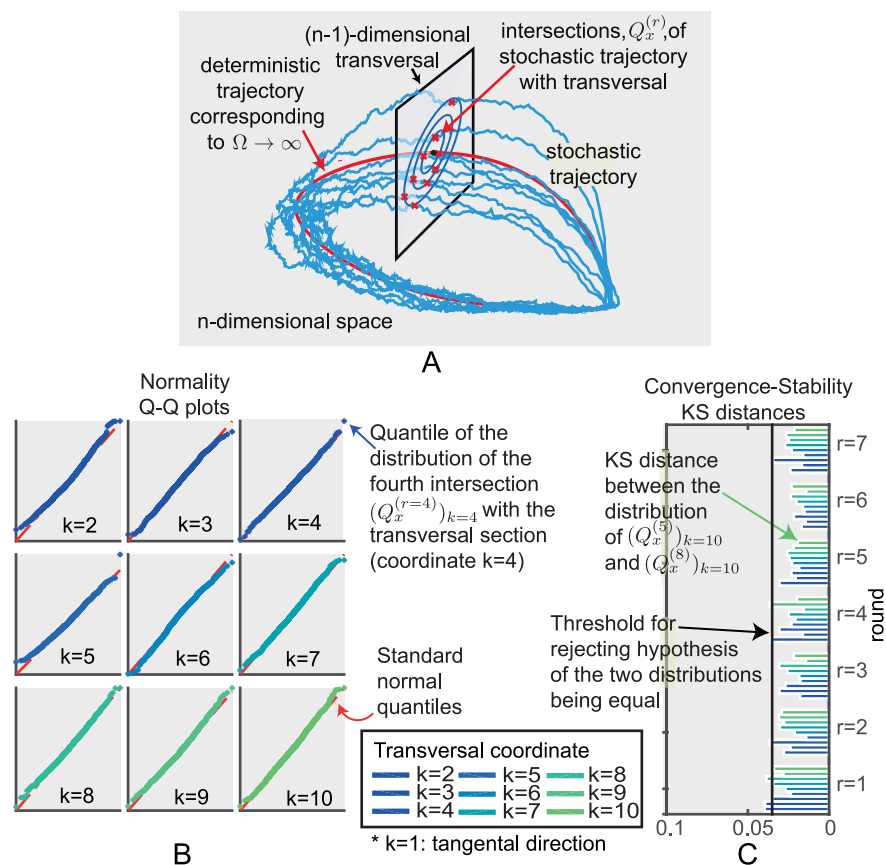


Fig 2. Intersections of the stochastic trajectories with a transversal section $\mathcal{S}_{g(t_0)}$. (A) Schematic representation of the intersections. (B) We consider an adapted coordinate system (x_1, \dots, x_{10}) (see S1 Sect. 1) so that (x_2, \dots, x_{10}) are orthogonal coordinates on $\mathcal{S}_{g(t_0)}$ and then consider the distribution $P^{k,r}$ of the values of x_k for $k=2, \dots, 10$ at the intersection points $Q_{g(t_0)}^{(r)}$. Quantile-Quantile (Q-Q) plots of these distributions for the fourth pass, $Q_{g(t_0)}^{(r=4)}$, show that the distributions are very close to being normal, (see Fig C in S2 Appendix for similar plots for $r=1, 2, \dots, 8$). (C) KS distances between the above distributions for the first $r=1, 2, \dots, 7$ rounds of the stochastic trajectory and the distribution in the 8th round ($r=8$), $k=2, 3, \dots, 10$.

<https://doi.org/10.1371/journal.pcbi.1005676.g002>

of intersection with the limit cycle. Therefore, these transversal distributions are useful for analysing various aspects of the system.

Our first observation is that the empirical transversal distributions

$$P_{x_1}^{(r)} = P(Q_{x_1}^{(r,X)} | X(t_0) = x_0)$$

obtained by exact simulation are approximately multivariate normal (Fig 2(B)). Moreover, as r increases $P_{x_1}^{(r)}$ and $P_{x_1}^{(r+1)}$ are hardly distinguishable and appear to converge to a fixed, approximately normal transversal distribution as r increases (Fig 2(C)). Similar results hold if we use a different family of transversal sections to γ as explained in S1 Sect. 8.2 & 8.3.

A natural question that arises is whether one obtains a different distribution when instead of taking the first intersection point of the r th pass one takes a later intersection point of the same round. This is addressed in S1 Sect. 12 where we show that in exact simulations there appears to be no difference as would be expected from the LNA approximation.

The Linear Noise Approximation (LNA)

The convergence of the transversal distributions to approximately normal distributions naturally raises the question of whether asymptotic approximation methods such as the LNA, which provide multivariate normally distributed approximation of the stochastic system, can be used to accurately approximate these transversal distributions.

The LNA as formulated by [6] is derived directly from the underlying Markov jump process and is valid for any time interval of finite fixed length. It is based on the ansatz

$$X(t) = \frac{Y(t)}{\Omega} = x(t) + \frac{\xi(t)}{\sqrt{\Omega}} \quad (3)$$

where $x(t)$ is a solution of the limiting ($\Omega \rightarrow \infty$) deterministic system Eq (1) and $\xi(t)/\sqrt{\Omega}$ describes the stochastic variations. In our case we always take $x(t)$ to be the periodic solution $g(t)$. A key aspect of this ansatz is that $\xi(t)$ satisfies a linear stochastic differential equation that is independent of Ω , with drift and diffusion matrices that are functions of the deterministic solution $g(t)$. Details are given in S1 Sect. 4 & 5.

Given an initial time t_0 and an initial condition $\xi(t_0)$ for ξ , the LNA determines the distribution, of $\xi(t)$, $t > t_0$, and hence $X(t) = g(t) + \xi(t)/\sqrt{\Omega}$ that we respectively denote by $P_{\text{LNA}}(\xi(t)|t_0, \xi(t_0))$ and $P_{\text{LNA}}(X(t)|t_0, \xi(t_0))$. If $\xi(t_0)$ is only known up to a multivariate normal (MVN) distribution P_0 then we denote these distributions, respectively, by $P_{\text{LNA}}(\xi(t)|t_0, \xi(t_0) \sim P_0)$ and $P_{\text{LNA}}(X(t)|t_0, \xi(t_0) \sim P_0)$. Details of how to calculate these distributions are given in S1 Sect. 4. Each of the above distributions is MVN enabling analytical approaches, for example in analysing the stochastic sensitivities of the system.

If we fix $t > t_0$ then as $\Omega \rightarrow \infty$ the true distribution of ξ converges to the distribution $P_{\text{LNA}}(\xi(t)|t_0, \xi(t_0))$ (see e.g. [5]). However, one most certainly cannot reverse the limits i.e. for a fixed Ω one cannot expect the approximation to hold for large time $t \rightarrow \infty$. As we now show, this is certainly the case for oscillators and we aim to overcome this limitation by developing methods that remain accurate for much larger times than the LNA.

We first consider the distribution $P(X(t)|X(0) = x_0)$ and compare this for SSA simulated samples and the LNA at a sequence of times $t = \tau, 2\tau, \dots, 8\tau$ and for an arbitrary (fixed) initial state $x_0 \in \gamma$. As we can see in Fig 3, the LNA fits the SSA simulations relatively well in the short run ($t \leq \tau$), but as time progresses the Kolmogorov-Smirnov (KS) distance between the two distributions for each state variable for the LNA and the SSA increases substantially beyond the threshold level (see Fig 3(B)). The LNA predictions spread along the tangential direction

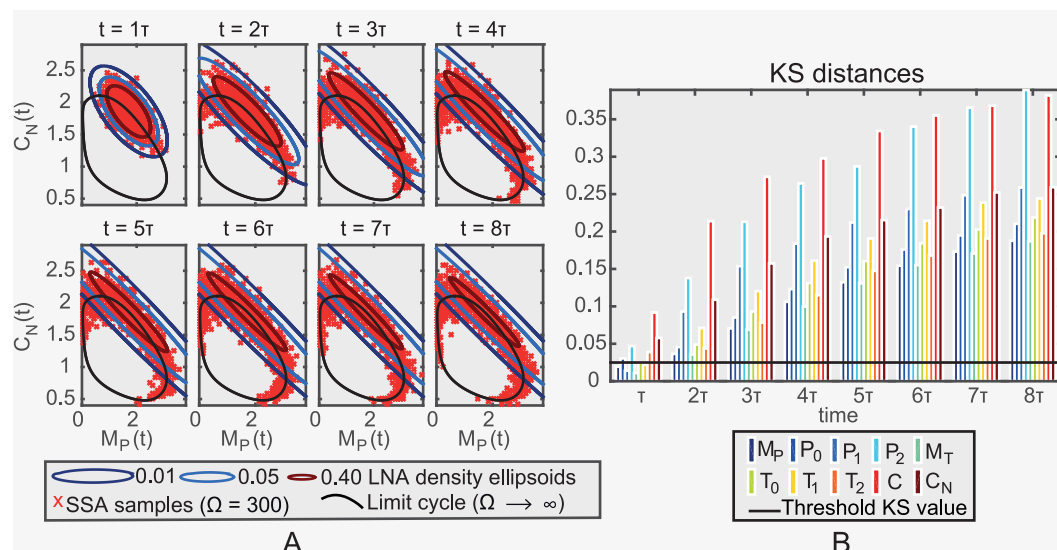


Fig 3. Comparison between LNA and exact simulations. (a) Samples (in nanomolar concentrations) obtained from the SSA simulation algorithm (red crosses) and 0.01, 0.05, 0.40 contours of the LNA probability density (colored ellipsoids) at fixed times, $t = \tau, 2\tau, \dots, 8\tau$ (τ : minimal period), for the circadian clock system (M_P per mRNA; C_N nuclear PER:TIM complex P_0 & T_0 PER & TIM protein; P_1 & T_1 phosphorylated PER & TIM protein; P_2 & T_2 twice phosphorylated PER & TIM protein; C cytoplasmic PER:TIM complex, units as in Fig 1). The limit cycle ODE solution is also displayed (black solid line). (b) KS distance between the empirical distribution of SSA samples and the LNA distribution of each species (different colors, see legend) at the fixed times. The threshold level is also displayed (black solid line). The system size is $\Omega = 300$.

<https://doi.org/10.1371/journal.pcbi.1005676.g003>

and therefore fail to accurately reflect the SSA samples that have instead spread along the curved limit cycle.

On the other hand, as we saw earlier, the transversal distributions $P_{x_1}^{(r)} = P(Q_{x_1}^{(r)} | X(t_0) = x_0)$ of the *Drosophila* circadian clock system are approximately normal (Fig 3(B)). We next derive an approximation of $P_{x_1}^{(r)}$ under the LNA and show that it accurately approximates $P_{x_1}^{(r)}$ for the *Drosophila* circadian clock, Brusselator and NF- κ B systems.

Calculating transversal distributions

We now generalise slightly and consider the set of stochastic trajectories X where the initial conditions $X(t_0)$ have a MVN distribution P_0 that is supported on the normal transversal section $\mathcal{S}_{g(t_0)}$ (denoted $X(t_0) \sim P_0$). We consider how to approximate the distribution $P_{x_1}^{(r, P_0)}$ of the intersection points $Q_{g(t_1)}^{(r, X)}$ of these trajectories with the normal transversal section $\mathcal{S}_{g(t_1)}$, $t_1 > t_0$. As an approximation we take the conditional distribution

$$P_{LNA, t_1}^{(r)} = P_{LNA}(X(t_1^{(r)}) | X(t_1^{(r)}) \in \mathcal{S}_{g(t_1)}, \zeta(t_0) \sim P_0), \quad t_1^{(r)} = t_1 + (r-1)\tau \quad (4)$$

given by conditioning $P_{LNA, t_1}^{(r, \text{free})} = P_{LNA}(X(t_1^{(r)}) | \zeta(t_0) \sim P_0)$ on $X(t_1^{(r)}) \in \mathcal{S}_{g(t_1)}$. It gives a MVN distribution supported on $\mathcal{S}_{g(t_1)}$.

In S1 we show that, although, for free-running oscillators, $P_{LNA, t_1}^{(r, \text{free})}$ diverges as $r \rightarrow \infty$, the mean and covariance of the MVN transversal distribution $P_{LNA, x_1}^{(r)}$ converge exponentially fast to those of a MVN distribution $P_{LNA, x_1}^{(\infty)}$ (S1 Sect. 8, cf. Fig 4). The distribution $P_{LNA, x_1}^{(\infty)}$ is a fixed point in the sense that if the distribution of $X(t_1 + \tau)$ is given by the LNA using as initial

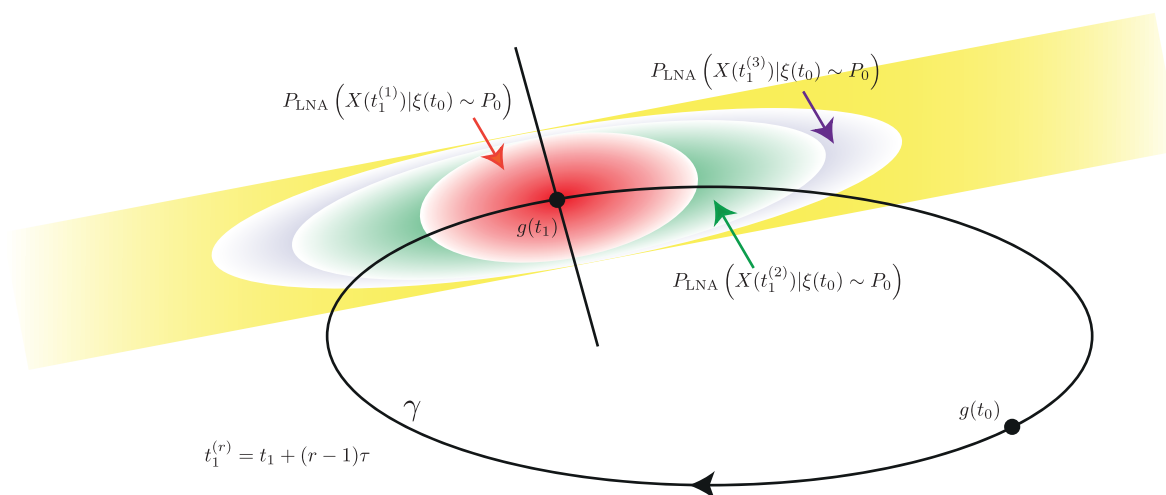


Fig 4. Schematic diagram of the distributions $P_{\text{LNA}}(X(t_1^{(r)}) | \xi(t_0) \sim P_0)$ as r increases. The yellow distribution represents the limit as $r \rightarrow \infty$. Although, for free oscillators the latter distribution diverges, the corresponding conditional distributions on the transversal converge.

<https://doi.org/10.1371/journal.pcbi.1005676.g004>

condition $\xi(t_1) \sim P_{\text{LNA},x_1}^{(\infty)}$ then conditioning on $X(t_1 + \tau) \in \mathcal{S}_{g(t_1)}$ gives

$$(X(t_1 + \tau) | X(t_1 + \tau) \in \mathcal{S}_{g(t_1)}) \sim P_{\text{LNA},x_1}^{(\infty)}.$$

Using this fact enables us to calculate $P_{\text{LNA},x_1}^{(\infty)}$ directly because we show in S1 that its mean and covariance matrix satisfy a simple fixed point equation that is easily solved numerically (S1 Sect. 9.1).

The reader will note that in Eq (4) we approximate by conditioning on $X(t_1^{(r)}) \in \mathcal{S}_{g(t_1)}$ whereas we should have conditioned on $X(t) \in \mathcal{S}_{g(t_1)}$ for arbitrary t corresponding to the r th round. In S1 Sect. 7 we argue that the error in the mean and variance of the distribution due to taking $t = t_1^{(r)}$ is $O(\Omega^{-1})$.

The question remains as to how well these distributions capture the exact simulation transversal distribution $P_{x_1}^{(r)}$. This is addressed in Fig 5 where it is shown that the fit is excellent even for Ω as low as 300. The fit is even better for higher system sizes (Fig E in S2 Appendix). In S3 we also show similar low Ω results for the Brusselator (Fig B in S3 Appendix) and the NF- κ B system (Fig E in S3 Appendix). The result is also true for the light-entrained *Drosophila* circadian clock system (see Fig J in S2 Appendix) and the transient oscillations of the NF- κ B system (see Fig H in S3 Appendix). Thus we note that although the LNA cannot be used directly to accurately compute $P(X(t) | X(0))$ for a fixed Ω and increasing t , using it to compute the transversal distributions provides accurate estimates of $P_{x_1}^{(r)}$ for much larger times $t_1 + r\tau$. Moreover, in S1 Sects. 8.2 & 8.3 we also explain why the convergence of the distribution on normal hyperplanes implies convergence on other transversal sections to γ .

In S1 Sect. 8.4, we explain that in contradistinction to free-running oscillators, for entrained forced oscillators, $P_r = P_{\text{LNA}}(X((r-1)\tau + t_1) | x_0, \xi_0 \sim P_0)$ converges as $r \rightarrow \infty$ so that, under the LNA, the phase fluctuations have a variance that is bounded independently of r . The corresponding conditional distribution is therefore a correspondingly good approximation to the transversal distribution $P_{x_1}^{(r,P_0)}$ (see Fig J in S2 Appendix). However, it does not mean that $P_{\text{LNA},t_1}^{(r,\text{free})}$ is a good approximation to the corresponding distribution $P(X(t_1^{(r)}) | X(0))$ for an exact

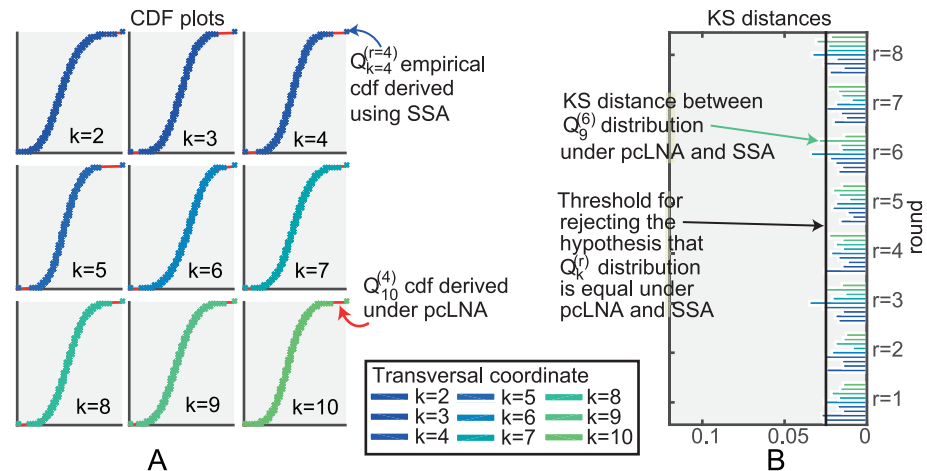


Fig 5. Comparison of pcLNA and exact transversal distributions for the *Drosophila* circadian clock ($\Omega = 300$). (A) CDF plots of the distributions $P^{k,r}$ as in Fig 2 for the pcLNA (red line) and the SSA (empirical CDF, crosses) for $k = 2, 3, \dots, 10$ and round $r = 4$ (see Fig D in S2 Appendix for $r = 1, 2, \dots, 8$). (B) KS distances between the corresponding distributions under pcLNA and SSA, $r = 1, 2, \dots, 8$, $k = 2, 3, \dots, 10$.

<https://doi.org/10.1371/journal.pcbi.1005676.g005>

simulation. In fact, we show in Fig I in S2 Appendix that $P_{\text{LNA}, t_1}^{(r, \text{free})}$ is a poor approximation of the empirical distribution $P(X(t_1^{(r)})|X(0))$ derived from exact simulations for the light-entrained *Drosophila* circadian clock ($\Omega = 300$). The bounded variance of the phase fluctuations as $r \rightarrow \infty$ for forced oscillators is the basic mechanism behind the population-level entrainment of stochastic oscillators introduced in [21].

Stochastic fluctuations in periods and timing

We now consider the fluctuations δt in the time taken for the lifted phase of a stochastic trajectory to go from a given phase φ_1 to a greater one φ_2 . If $\varphi_2 - \varphi_1 = 2r\pi$ then this corresponds to the time taken to perform r cycles.

In Fig 6 we give an example using the *Drosophila* circadian clock model where we take φ_1 to be 0 (with a fixed initial condition $x_0 \in \gamma$) and $\varphi_2 = 2r\pi$ for $r = 1, \dots, 8$. The distributions of δt appear to be very close to normal and the variance appears to grow linearly with r . We also consider the case where $\varphi_1 = 2(r-1)\pi$ and $\varphi_2 = 2r\pi$ for $r = 1, \dots, 8$. Again the distributions are approximately normal but the variances are approximately constant (Fig 6(C)). Because for a given r the trajectory has done $r-1$ cycles before reaching the lifted phase φ_1 the distribution of the state at this phase is changing with r . We expect that this distribution is converging with increasing r and this result (Fig 6(C)) is in accordance with this.

In S1 Sect. 6 we approximate the statistics of δt using the LNA and show that as a random variable δt is approximately normal with mean that is $O(\Omega^{-3/2})$ and we also calculate its variance up to terms that are $O(\Omega^{-3/2})$ and the extent of its divergence from normality.

All points with a given lifted phase φ_1 lie in a particular transversal $\mathcal{S}_{g(t_1)}$ with $0 \leq t_1 < \tau$. If $t_2 = t_1 + (\varphi_2 - \varphi_1)\tau/2\pi$, then, the mean and variance of δt can be calculated in terms of t_1 and t_2 . If the initial conditions $\xi(t_1)$ are MVN distributed on $\mathcal{S}_{g(t_1)}$ with mean 0 and covariance V_1 , this variance is $(1/\alpha^2\Omega) \check{V}_{11} + O(\Omega^{-3/2})$ where $\check{V}_{11} = \check{V}(t_1, t_2)_{11}$,

$$\check{V}(t_1, t_2) = C(t_1, t_2) V_1 C(t_1, t_2)^T + V(t_1, t_2).$$

written in adapted coordinates at $g(t_2)$ (see S1 Sect. 1). All terms on the right hand side of this

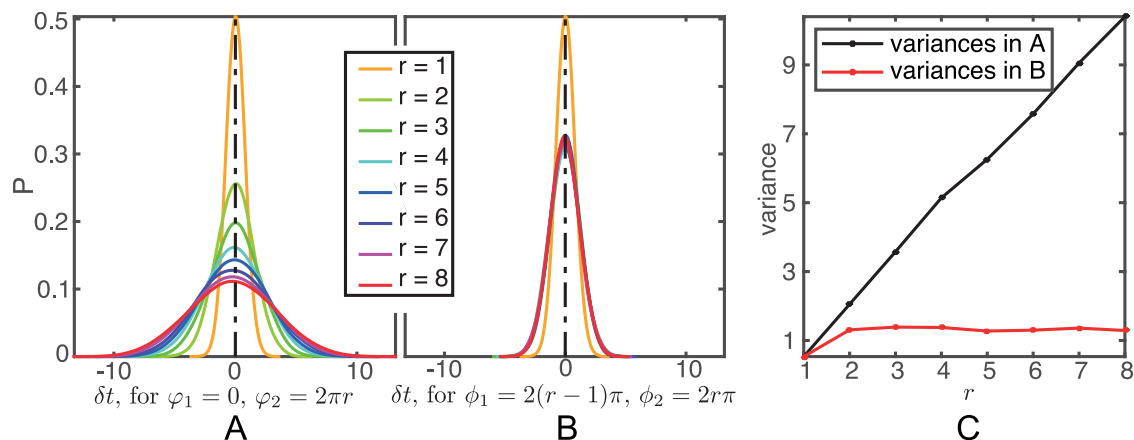


Fig 6. Exact empirical distribution of the fluctuations δt in *Drosophila* circadian clock system size $\Omega = 300$. (A) The (smoothed) empirical density function of δt in the time taken for the lifted phase of a stochastic trajectory to go from $\phi_1 = 0$ to $\phi_2 = 2\pi r$, $r = 1, 2, \dots, 8$. (B) The (smoothed) empirical density function of the fluctuations in the time taken for the lifted phase of a stochastic trajectory to go from $\phi_1 = 2(r-1)\pi$ to $\phi_2 = 2\pi r$, $r = 1, 2, \dots, 8$. (C) The variance of the distributions in (A) and (B).

<https://doi.org/10.1371/journal.pcbi.1005676.g006>

equation are defined in S1 Sect. 4. The above exact simulations of the *Drosophila* circadian clock agree with these theoretical predictions. It is easy to see (cf. S1 Sect. 8) that \tilde{V}_{11} grows roughly linearly with $t_2 - t_1$.

pcLNA simulation algorithm

Given the ability to accurately approximate the transversal distributions and the results in [9] we realised it should be possible to use this to construct a rapid simulation algorithm. The linear increase of the variance of the deviations δt , or equivalently, the linear growth in the variance of the deviation of the lifted phase $\bar{\varphi}_X(t_i)$ from $2\pi t_i/\tau$, indicates the reasons for the long-time failure of the standard LNA. It is unable to cope with the increasing phase deviations. This motivates the phase correction approach used in the simulation algorithm we now define.

The approach is to amend the LNA Ansatz $X(t) = g(t) + \Omega^{-1/2} \xi(t)$ to $X(t) = g(s) + \Omega^{-1/2} \kappa(s)$ where $g(s) = G_N(X(t))$ and to use resetting of t to s to cope with the growth in the variance of $\bar{\varphi}_X(t_i) - 2\pi t_i/\tau$ keeping the LNA fluctuation $\kappa(s)$ normal to γ . While for free-running oscillators the variance of $\xi(t)$ grows without bound as t increases, $\kappa(s)$ has uniformly bounded variance.

The pcLNA simulation algorithm iteratively uses standard LNA steps of length $\Delta\tau$ to move from a state $X(s_{i-1})$ to a new state $X(s_{i-1} + \Delta\tau) = X_i$, $i = 1, 2, \dots$. After each LNA step, the phase of the system is reset or “corrected” such that $g(s_i) = G_N(X_i)$ and the (global) fluctuations $\xi(s_{i-1} + \Delta\tau) = \Omega^{1/2}(X_i - g(s_{i-1} + \Delta\tau))$ are replaced by the normally transversal fluctuation $\kappa(s_i) = \Omega^{1/2}(X_i - g(s_i))$ which are MVN distributed and, as we showed in the previous section, approximate well the transversal fluctuations under the exact Markov Jump process.

The steps of the pcLNA simulation algorithm are described next in more detail (see also Fig 7).

1. Choose a time-step size $\Delta\tau > 0$.
2. Input **initial conditions** $\kappa(s_0)$ and $X_0 = g(s_0) + \Omega^{-1/2} \kappa(s_0)$.
3. **For iteration** $i = 1, 2, \dots$

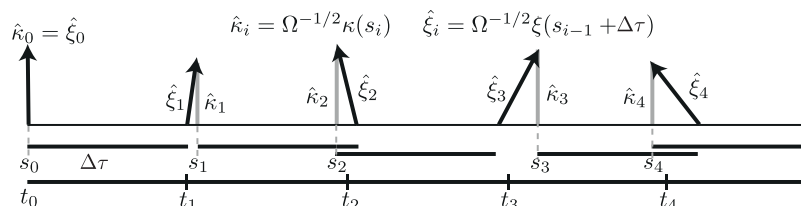


Fig 7. The main step in the pcLNA algorithm. The solid horizontal bars below the horizontal axis are all of length $\Delta\tau$, the basic time step of the algorithm. The black arrows show $\hat{\xi}_i = \Omega^{-1/2}\xi(s_{i-1} + \Delta\tau)$ and the grey arrows $\hat{\kappa}_i = \Omega^{-1/2}\kappa(s_i)$. Having calculated $\hat{\kappa}(s_{i-1})$ one uses $\kappa(s_{i-1})$ as the initial state and updates it using the LNA and a time-step $\Delta\tau$ to obtain ξ at $s_{i-1} + \Delta\tau$. Then $\xi(s_{i-1} + \Delta\tau)$ is replaced by $\kappa(s_i)$ so that $g(s_{i-1} + \Delta\tau) + \Omega^{-1/2}\xi(s_{i-1} + \Delta\tau) = g(s_i) + \Omega^{-1/2}\kappa(s_i)$ where $\kappa(s_i)$ is normal to the limit cycle. Therefore, s_i gives the phase of $\kappa(s_i)$ and the corresponding time is $t_i = t_0 + i\Delta\tau$.

<https://doi.org/10.1371/journal.pcbi.1005676.g007>

- sample $\xi(s_{i-1} + \Delta\tau)$ from $\text{MVN}(C_i \kappa(s_{i-1}), V_i)$;
- compute $X_i = g(s_{i-1} + \Delta\tau) + \Omega^{-1/2} \xi(s_{i-1} + \Delta\tau)$;
- set s_i to be such that $G_N(X_i) = g(s_i)$ and $\kappa_i = \Omega^{1/2}(X_i - g(s_i))$.

In the for loop $C_i = C(s_{i-1}, s_{i-1} + \Delta\tau)$ and $V_i = V(s_{i-1}, s_{i-1} + \Delta\tau)$ are the drift and diffusion matrices in the linear SDE describing the evolution of the noise process $\xi(t)$ under the LNA (see S1 Sect. 4).

The simulated sample X_i corresponds to time $t_i = t_0 + i\Delta\tau$, $i = 1, 2, \dots$, where t_0 is the initial time. The time t_i is not necessarily equal to the phase s_i , defined by the relation $g(s_i) = G_N(X_i)$, which is stochastic and has variance linearly increasing with the time step $\Delta\tau$.

If one wants to record simulated trajectories at a finer time-scale than $\Delta\tau$ then one can run the algorithm with $\Delta\tau$ replaced by $\Delta\tau/M$ for some integer $M > 1$ and only carry out the phase correction in step 3(c) every M th step and at all the other steps just proceeding as in the standard LNA (ignoring step 3(c)). This gives the same distribution as if the intermediate points had not been calculated because of the transitive nature of the LNA i.e. the distribution $P_{\text{LNA}}(X(s+t)|X(0))$ is equal to the distribution $P_{\text{LNA}}(X(t)|X(s) \sim P_{\text{LNA}}(X(s)|X(0))$. In the simulation results described below the time-step $\Delta\tau = 6$ hours and $M = 3$ so that there are $\tau/6 \approx 4.5$ corrections in every round of the limit cycle. The effect of less frequent correction is studied in S2 Sect. 5.

Comparisons to other simulation algorithms. We compared the pcLNA simulation algorithm in terms of both CPU time and precision of the approximation with some of the most important alternatives, the tau-leap method and integration of the chemical Langevin equation (CLE) method. Exact simulations produced by the SSA are also used as a reference for the comparison.

For the tau-leap simulation we use the algorithm described in [3], which is a refinement of the original tau-leap method first proposed in [2]. In [3] the authors suggest an optimal method to compute the largest possible time step such that the leap condition is still approximately satisfied. The leap condition ensures that the state change in any time step is small enough so that no rate function will experience a macroscopically significant change in its value. The error of this approximation is controlled by a parameter ϵ . For integrating the CLE described in [4], we use the classical Euler method (see [32]) with a fixed time step Δt . The integration of the CLE can be done using methods that include higher-order terms in the integration and this has been shown to improve the speed of implementation in low-dimensional systems, albeit with a cost in the complexity of the algorithm (see [33, 34]). However, we are not aware of any implementations of these methods for high-dimensional systems such as

those considered here, or of comparisons of such methods to the Euler method for such problems.

For both the tau-leap and the CLE approximation we explored different values of their parameters ϵ and Δt to attain a good balance of precision and CPU time. Here we present the results for the largest values of both ϵ and Δt , hence smallest CPU time, which attain similar performance with the pcLNA algorithm in terms of precision. If little improvement could be achieved in terms of precision by lowering either ϵ or Δt , the larger values are preferred.

The algorithms are implemented for a fixed time-interval (8.5 times the period of the limit cycle of the system) and the comparison is made at 8 fixed time points using the KS distances between the empirical distribution of each algorithm and the empirical distribution derived using the SSA simulations. Note that for all approximation algorithms considered here, the probability of generating negative populations is non-zero and there are a number of methods for dealing with this. Our simple approach is described in S1 Sect. 13.

Fig 8 displays the median CPU times for a single trajectory simulation in $t \in [0, 8.5\tau]$, under the competing approaches for different system sizes, $\Omega = 300, 1000, 3000$, along with a comparison in terms of precision for $\Omega = 300$ (see Figs G & H in S2 Appendix for $\Omega = 1000$ & 3000). A sample of size $R = 2000$ is produced for each algorithm. We see that the precision of all approximation methods is fairly similar, with their empirical distributions almost indistinguishable to exact simulations. In terms of CPU times, we see that the SSA is much slower than the other algorithms particularly for large system sizes. The tau-leap offers some improvement to the CPU times but this is relatively small compared to the CLE approximation and especially the pcLNA algorithm. One reason for this relatively small improvement for the tau-leap algorithm is the stiffness of the considered system, a property that is however very common in biological systems and it is known to slow down the tau-leap method by requiring small values of the ϵ parameter to ensure that the leap condition is satisfied and hence the approximation is fairly precise. Note that for similar reasons, a small Δt was necessary to achieve good precision with the CLE approximation.

As we can see in Fig 8, the pcLNA algorithm is about 24 times faster than the CLE approximation, tens to hundreds of times faster than the tau-leap and hundreds to thousands of times faster than the SSA. In our simulations, it took 0.4sec to derive this long-time trajectory, which means that in about 7 minutes one can generate more than 1000 trajectories of this large system over a long-time compared to about 2.7 hours with CLE approximation and much longer times for the other methods. Therefore, the pcLNA offers a substantial improvement in CPU times compared to standard approaches in simulating oscillatory systems without compromising the precision of the simulation substantially.

Perhaps more importantly, this pcLNA simulation algorithm has the advantage of being based on an analytical framework that allows calculation of some key distributions. Therefore, it enables more rigorous methods for assessing accuracy and robustness.

Analysis and inference of oscillatory systems using pcLNA

The derivation of analytical expressions of the transversal distributions allows us to analyse various aspects of the stochastic behaviour of these systems that can possibly involve a large number of variables and parameters. Here we illustrate the use of pcLNA transversal distributions to perform such an analysis. We begin by describing the pcLNA joint distribution of multiple intersections to possibly different transversal sections on the limit cycle and then discuss Fisher information, sensitivity analysis and estimation by Kalman filtering.

Joint distribution on multiple transversals. Consider q phase states of the limit cycle $x_i = g(t_i)$, $i = 1, \dots, q$, on γ where $0 \leq t_1 < t_2 < \dots < t_q < \tau$. If $X(t)$ is a stochastic trajectory, we

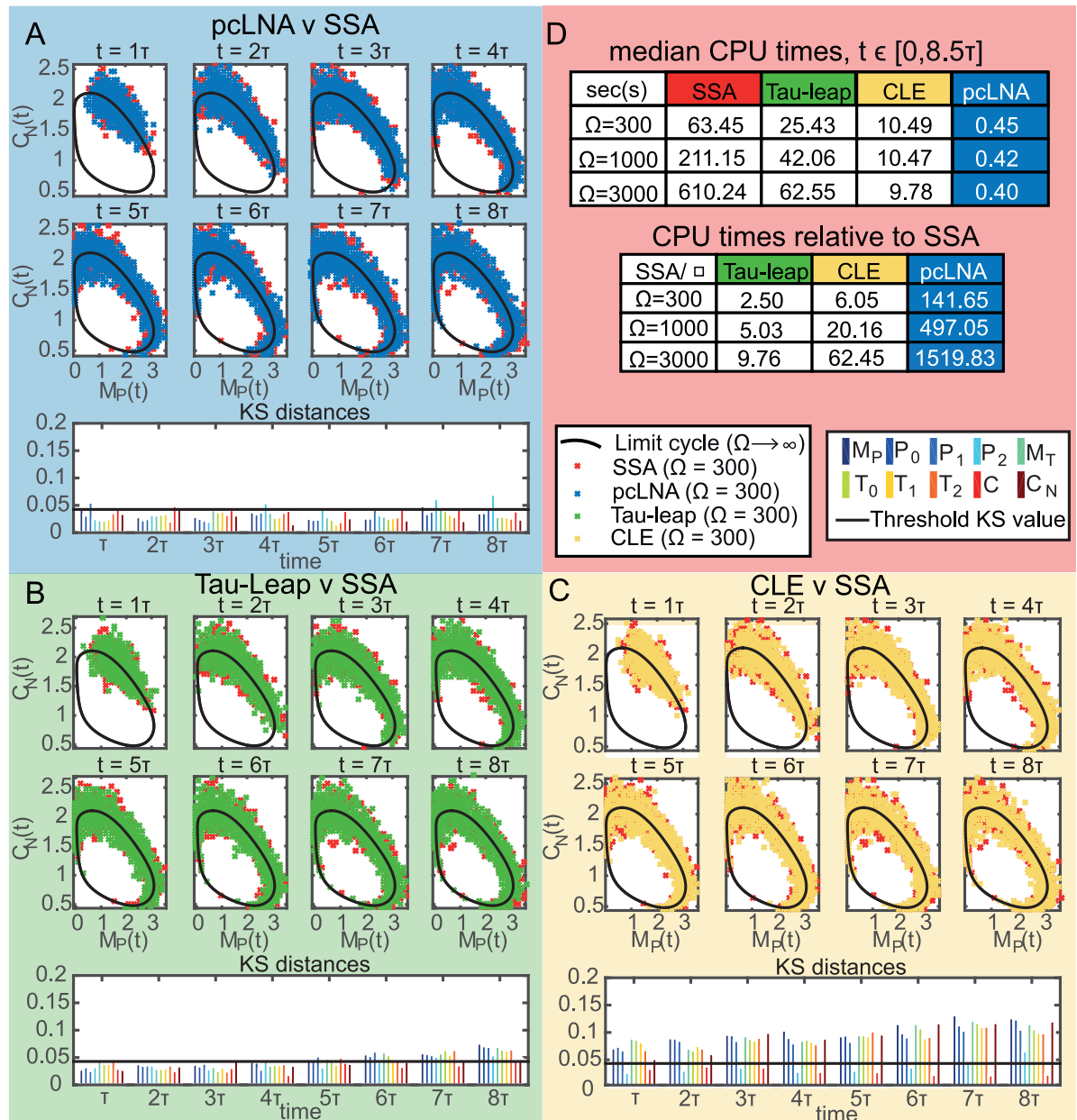


Fig 8. Comparison between pcLNA, tau-leap and CLE simulation algorithms for the *Drosophila* circadian clock. Panels (A), (B) and (C) contain the samples (in concentrations, units as in Fig 1) produced by respectively the pcLNA, tau-leap ($\epsilon = 0.002$) and CLE ($\Delta t = 0.002$) algorithms and the exact simulation (SSA) at time-points $t = \tau, 2\tau, \dots, 8\tau$, for $\Omega = 300$, along with the KS distances between the empirical distributions of each approximation and the SSA for each system variable (coloured bars; variable names as in Fig 3). Panel (D) provides the median CPU times for a single trajectory run in the time-interval $[0, 8.5\tau]$ under the different simulation algorithms along with the ratio of the median CPU time under SSA and each approximation algorithm.

<https://doi.org/10.1371/journal.pcbi.1005676.g008>

consider how it meets the transversal sections at the x_i as t increases using the lifted phase function $\bar{\varphi}_X$. We can talk of the times when $G_N(X(t))$ first takes the phase x_i during the r -th revolution of $G_N(X(t))$ around γ . Using $\bar{\varphi}_X$ to define the points $Q_{g(t)}^{(r)}$ we have that it first meets \mathcal{S}_{x_i} in $Q_{x_i}^{(1)}$ for $i = 1, \dots, q$. If $i < q$ then we let $Q_{x_{i+1}}^{(k)}$ denote the first point in $\mathcal{S}_{x_{i+1}}$ that X meets after it leaves $Q_{x_i}^{(k)}$. If $i = q$ then the next transversal it meets is \mathcal{S}_{x_1} and the intersection point is $Q_{x_1}^{(k+1)}$.

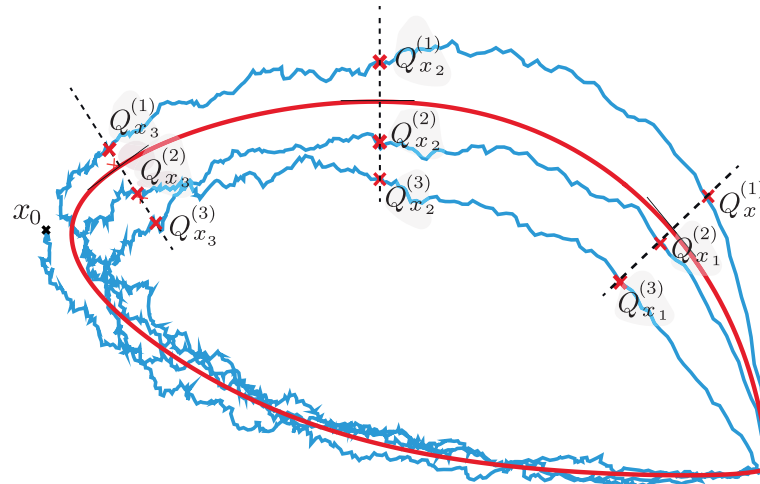


Fig 9. The sequence Q in two-dimensions. The stochastic trajectory $X(t)$ (blue line), initiated from x_0 , intersects each of the transversal sections x_1 , x_2 and x_3 (dashed lines) of the limit cycle (red solid line) three times, following a path $Q_{x_1}^{(1)} \rightarrow Q_{x_2}^{(1)} \rightarrow Q_{x_3}^{(1)} \rightarrow Q_{x_1}^{(2)} \rightarrow Q_{x_2}^{(2)} \rightarrow Q_{x_3}^{(2)} \rightarrow Q_{x_1}^{(3)} \rightarrow Q_{x_2}^{(3)} \rightarrow Q_{x_3}^{(3)}$.

<https://doi.org/10.1371/journal.pcbi.1005676.g009>

In this way (see Fig 9) we derive a sequence of intersection points $\underline{Q} = Q_{x_1}^{(1)}, \dots, Q_{x_q}^{(1)}, Q_{x_1}^{(2)}, \dots, Q_{x_q}^{(2)}, \dots, Q_{x_1}^{(m)}, \dots, Q_{x_q}^{(m)}$.

We shall be interested in the distribution

$$P(\underline{Q} | X(t_0)) = P(Q_{x_1}^{(1)}, \dots, Q_{x_q}^{(m)} | X(t_0)). \quad (5)$$

Remarkably, in our approximation, this distribution is MVN with a covariance matrix whose inverse has a simple tridiagonal form in terms of the drift and diffusion matrices coming from the LNA (S1 Sects. 9.2 & 9.3).

The fact that the above transversal distributions are MVN allows us to analytically compute the Fisher Information matrix and associated quantities that can be used to perform a stochastic sensitivity analysis of oscillatory systems.

Fisher information

Fisher Information quantifies the information that an observable random variable carries about an unknown parameter θ . If $P(X, \theta)$ is a probability distribution depending on parameters θ , the Fisher Information Matrix (FIM) $I = I_P$ has entries

$$I_{ij} = E \left[\frac{\partial \ell}{\partial \theta_i} \frac{\partial \ell}{\partial \theta_j} \right] = -E \left[\frac{\partial^2 \ell(\theta; X)}{\partial \theta_i \partial \theta_j} \right], \quad (6)$$

where $\ell = \log P$, and θ_i and θ_j are the i th and j th components of the parameter θ . If P is MVN with mean and covariance $\mu = \mu(\theta)$ and $\Sigma = \Sigma(\theta)$ then

$$I_{ij} = \frac{\partial \mu^T}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} \text{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right). \quad (7)$$

The FIM measures the sensitivity of P to a change in parameters in the sense that

$$D_{KL}(P(\cdot, \theta + \delta\theta), P(\cdot, \theta)) = \frac{1}{2} \delta\theta^T I_P \delta\theta + O(\|\delta\theta\|^3)$$

where D_{KL} is the Kullback-Leibler divergence. The significance of the FIM for sensitivity and experimental design follows from its role in Eq (6) as an approximation to the Hessian of the log-likelihood function at a maximum. Assuming non-degeneracy, if θ^* is a parameter value of maximum likelihood there is a $s \times s$ orthogonal matrix V such that, in the new parameters $\theta' = V \cdot (\theta - \theta^*)$,

$$\ell(\theta) \approx \ell(\theta^*) - \sum_i \sigma_i^2 \theta_i'^2.$$

for θ near θ^* . From these facts it follows that the σ_i^2 are the eigenvalues of the FIM and that the matrix V diagonalises it. If we assume that the σ_i are ordered so that $\sigma_1^2 \geq \dots \geq \sigma_s^2$ then it follows that near the maximum the likelihood is most sensitive when θ_1' is varied and least sensitive when θ_s' is. Moreover, σ_i is a measure of this sensitivity.

The theory of optimal experimental design is based on the idea of trying to make the σ_i decrease as slowly as possible so that the likelihood is as peaked as possible around the maximum, thus maximising the information content of the experimental sampling methods. Various criteria for experimental design have been proposed including D-optimality that maximises the determinant of the FIM and A-optimality that minimises the trace of the inverse of the FIM [6]. Diagonal elements of the inverse of FIM constitute a lower-bound for variance of any unbiased estimator of elements of θ (Cramer-Rao inequality). However, for the systems we consider here the σ_i typically decrease very fast and there are many of them. Thus, in general, criteria based on a single number are more likely to be of less use than consideration of the set of σ_i as a whole.

Calculation of the FIM for stochastic systems using the LNA has been carried out in [22] but only for small systems and short times where the LNA is accurate. It is notable that the pcLNA enables one to do such sensitivity analysis for large systems and large times. As an example, we analyse the stochastic behaviour of the *Drosophila* circadian clock based on the limit distribution $P(\underline{Q} | Q_0)$ when

$$\underline{Q} = Q_{x_0}^{(1)}, Q_{x_1}^{(1)}, Q_{x_0}^{(2)}, Q_{x_1}^{(2)}, \dots, Q_{x_0}^{(m)}, Q_{x_1}^{(m)}$$

where $x_0 = g(t_0)$ and $x_1 = g(t_1)$ are chosen so that t_0 is the time of the peak of *per* mRNA M_p , and t_1 is the peak of the nuclear complex of PER and TIM proteins C_N . We compute the Fisher Information of the distribution $P(\underline{Q} | Q_0)$ using the closed form expression (S1 Sect. 9.3) for this distribution. As we can see in Fig 10(A) the eigenvalues of the Fisher Information matrix decay exponentially, with a sharp decline followed by a slower decrease. This reveals that the influential directions in the parameter space of the system are much less than its total dimension and that only a few parameters appear to be most influential. The eigenvectors associated with the two largest eigenvalues of the Fisher Information matrix (see Fig 10(B)) have large entries only for the parameters k_{dn} (PER-TIM complex nuclear degradation), k_d (*per* mRNA linear degradation), k_2 (PER-TIM complex transportation to cytosol), v_{st} (*tim* mRNA transcription), k_{ip} (*per* mRNA Hill coefficient) and k_{it} (*tim* mRNA Hill coefficient).

The exponential decrease of the eigenvalues is typical of tightly coupled deterministic systems [25–31], but has to our knowledge not been demonstrated before for stochastic systems. It has important consequences. For example, it tells us that only a few parameters can be estimated efficiently from time-series data unless the system is perturbed in some way to get complementary data and that there will be identifiability problems that can be analysed using the FIM. It can also be used to design experiments by considering the FIM of a combination of models including one for the proposed new experiment, choosing the new experiment so as to optimally alleviate the decline of the eigenvalues.

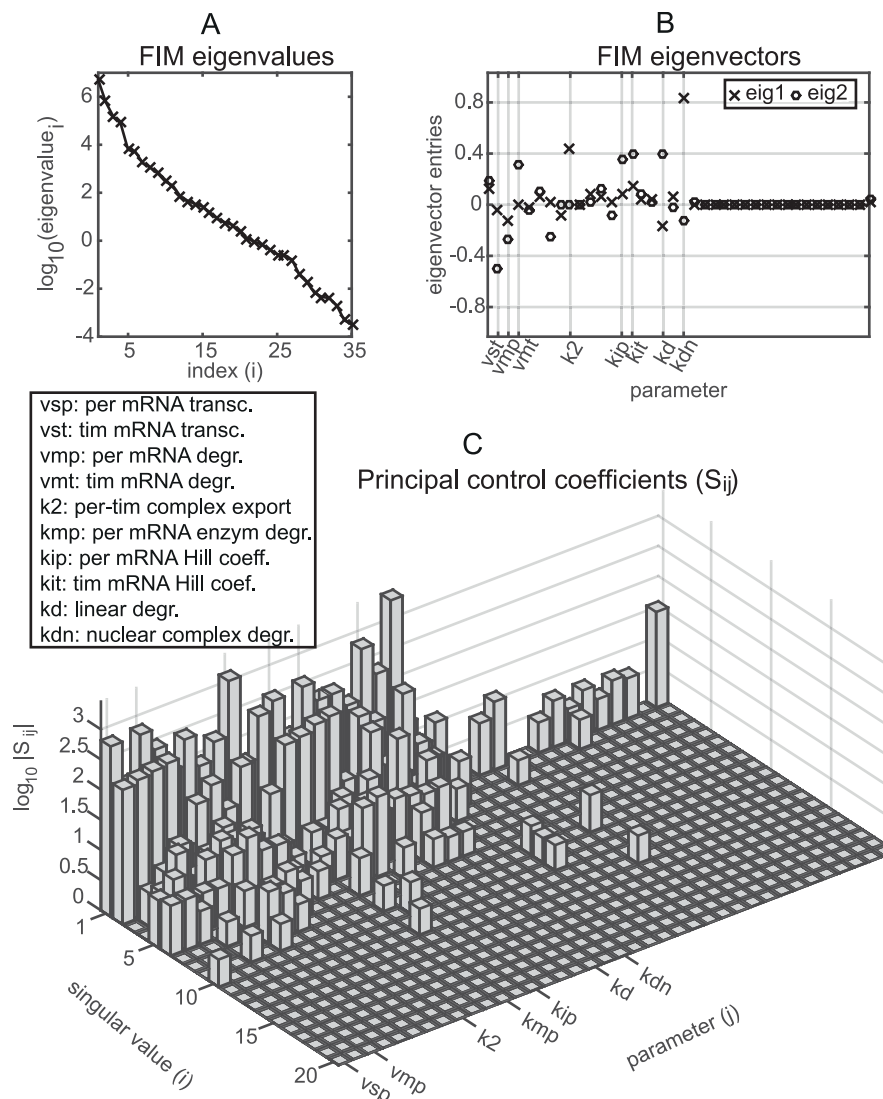


Fig 10. Fisher information and stochastic sensitivity analysis of the transversal distribution of the *Drosophila* circadian clock system at the times of the peaks of *per* mRNA and the nuclear complex of PER and TIM proteins. (A) The logarithm of the eigenvalues of the Fisher Information Matrix (FIM). (B) The entries/weights of the eigenvectors corresponding to the 2 largest eigenvalues of FIM (C) The largest principal control coefficients S_{ij} . Small values are reduced to 0.

<https://doi.org/10.1371/journal.pcbi.1005676.g010>

Sensitivity analysis for stochastic systems

The fact that we can calculate the Fisher Information allows a new approach to sensitivity analysis for stochastic systems. Anderson [23] and Srivastava et al. [24] also perform sensitivity analysis for small stochastic systems (up to 4 species and 8 reactions) in which they calculate the dependence of certain summary functions or statistics at one or more times to individual parameters. Our approach is different in that we use the fact that our distributions of interest are MVN and measure the change in the distribution of the system state at any given set of phases without recourse to any summary function and, moreover, this change is calculated for any combination of parameter variations. A major difference is our use of SVD below to find a basis of mean-covariance space using the principal components that enables us to decompose

these changes into different orthogonal directions that pick out the important and unimportant directions. The approach can also be formulated for the wider class of exponential families i.e. distributions that admit a representation of the form

$$P(x) = \exp \left\{ C(x) + \sum_i \theta_i F_i(x) - \varphi(\theta) \right\}$$

in terms of functions C, F_1, \dots, F_m of the state variable x and a function φ of the parameters θ .

We consider a family of probability distributions $P(X, \theta)$ which we assume are MVN with mean $\mu(\theta)$ and covariance matrix $\Sigma(\theta)$ depending on the parameters θ . We show that there is a natural matrix of sensitivities S_{ij} associated with such a system. These are system-global in that they look at how all components of the systems change with parameters. They also have an intimate relationship with Fisher information. Note that these results are not restricted to the transversal distributions derived in previous sections but apply more generally to any MVN distribution with mean $\mu = \mu(\theta)$ and covariance matrix $\Sigma = \Sigma(\theta)$ parameterised by a s -dimensional vector $\theta, s \geq 1$.

As is well-known in Information Geometry, the set of multivariate normal distributions MVN^n on \mathbb{R}^n can be given the structure of a Riemannian manifold in which the Riemannian metric is given by the line element

$$ds^2 = d\mu^T \Sigma^{-1} d\mu + (1/2) \text{tr}\{(\Sigma^{-1} d\Sigma)^2\}.$$

Points in MVN^n are denoted by $\Theta = (\mu, \Sigma)$ where μ is the mean and Σ the covariance matrix. The corresponding inner product in the tangent space at $\Theta_0 = (\mu, \Sigma)$ is given by

$$\langle \delta\Theta_1, \delta\Theta_2 \rangle_{\Theta_0} = \delta\mu_1^T \Sigma^{-1} \delta\mu_2 + \frac{1}{2} \text{tr}(\Sigma^{-1} \delta\Sigma_1 \Sigma^{-1} \delta\Sigma_2) \quad (8)$$

where $\delta\Theta_j = (\delta\mu_j, \delta\Sigma_j), j = 1, 2$.

In calculating the FIM we have to determine the partial derivatives $\partial\mu/\partial\theta_i$ and $\partial\Sigma/\partial\theta_i$. The derivative M of the mapping $\theta \rightarrow (\mu(\theta), \Sigma(\theta))$ at a parameter value θ_0 is given by

$$M \cdot \delta\theta = \left(\sum_i \frac{\partial\mu}{\partial\theta_i} \delta\theta_i, \sum_i \frac{\partial\Sigma}{\partial\theta_i} \delta\theta_i \right)$$

where the derivatives are calculated at θ_0 .

We can regard M as a linear mapping between the parameter space \mathbb{R}^s and MVN^n with the inner product given in Eq (8). We can then prove (S1 Sect. 11) that we can find s orthonormal vectors V_i spanning the parameter space \mathbb{R}^s , s orthonormal vectors U_i in the space MVN^n and numbers $\sigma_1 \geq \dots \geq \sigma_s \geq 0$ such that

$$MV_i = \sigma_i U_i, \quad i = 1, \dots, s. \quad (9)$$

Note that the orthonormality of the U_i is with respect to the inner product $\langle \cdot, \cdot \rangle_{\Theta_0}$. The eigenvalues of the FIM F are the squares of the σ_i because with respect to the standard inner product on θ -space and $\langle \cdot, \cdot \rangle_{\Theta_0}$ on MVN^n the adjoint M^* satisfies $M^* M = F$ (S1 Sect. 11).

If we let $U_i = (U_i^\mu, U_i^\Sigma)$ denote the decomposition of U_i into μ and Σ components, then the following key property follows from Eq (9): if $\delta\theta$ is any change of parameters, the change in μ

and Σ is given by

$$\begin{aligned}\delta\mu &= \sum_i U_i^\mu \left(\sum_j S_{ij} \delta\theta_j \right) + O(\|\delta\theta\|^2) \\ \delta\Sigma &= \sum_i U_i^\Sigma \left(\sum_j S_{ij} \delta\theta_j \right) + O(\|\delta\theta\|^2)\end{aligned}\quad (10)$$

where $S_{ij} = \sigma_i V_{ji}$.

One can define other sensitivities in a similar way but using a different orthogonal basis of MVN^n , but the above S_{ij} satisfy an important optimality condition explained in S1 Sect. 11 which asserts that the basis U_i and the corresponding sensitivities S_{ij} are optimal for capturing as much sensitivity as possible in the low order principal components U_i .

In view of this we call the S_{ij} the *principal control coefficients*. Note that the role of the S_{ij} as sensitivities is seen from the following relation which follows from Eq (10) (where $S = (S_{ij})$),

$$\|\delta\Theta\| = \|S \cdot \delta\theta\| + O(\|\delta\theta\|^2). \quad (11)$$

These sensitivities are relatively easy to calculate using the information in S1 Sect. 11. In Fig 10(C) we show the S_{ij} for the transversal distribution of the *Drosophila* circadian clock at the times of the peak of *per* mRNA and the peak of the nuclear complex of PER and TIM proteins. As we can see, because $S_{ij} = \sigma_i V_{ji}$ the coefficients rapidly decrease with the singular values σ_i , while a few parameters, similar to those with large eigenvector entries, have high coefficients.

Calculating likelihoods via a pcLNA Kalman Filter

The likelihood function of a set of time-series observations of a system can be used for parameter estimation, hypothesis testing and other forms of statistical inference. For example, one may wish to use the likelihood function to estimate parameters of a biological system. Although there is no elegant formula for

$$P(\underline{X} | X(t_0)) = P(X(t_1), \dots, X(t_m) | X(t_0))$$

similar to that for $P(\underline{Q} | X(t_0))$ above, we can efficiently calculate it. To do this we derive a Kalman Filter for the pcLNA that is a modification of the Kalman Filter associated with the LNA [35]. This can be used to compute the likelihood function $L(\theta; \hat{\underline{X}})$ of the system parameters θ with respect to observations $\hat{\underline{X}} = (\hat{X}(t_0), \hat{X}(t_1), \dots, \hat{X}(t_N))$ recorded at N times t_0, t_1, \dots, t_N that are noisy linear functions of the species concentrations. This is slightly more general than just calculating $P(\underline{X} | X(t_0))$ because we allow for a measurement equation. The Kalman filter can also be used for forward prediction.

We assume the measurement equation,

$$\hat{X}(t) = BX(t) + \epsilon, \quad (12)$$

relating the observations $\hat{X}(t)$ to the state variables, $X(t)$. Here B is a transformation matrix (often simply removing unobserved species or introducing unknown scalings) and $\epsilon = (\epsilon_1, \dots, \epsilon_n) \sim MVN(0, \Sigma_\epsilon)$ the observational error. The pcLNA likelihood can be decomposed as the product

$$L(\theta; \hat{\underline{X}}) = P(\hat{X}(t_0); \theta) \prod_{i=1}^n P(\hat{X}(t_i) | \hat{X}(t_{i-1}); \theta).$$

The pcLNA Kalman Filter algorithm, which we describe in more detail in S1 Sect. 15, uses a recursive algorithm for computing the terms in $L(\theta; \hat{X})$. The algorithm proceeds by iteration $i = 1, 2, \dots$ and uses Bayes rule to derive the posterior distributions of $(X(t_{i-1})|\hat{X}(t_{i-1}))$ and a phase correction to obtain $g(s_{i-1}) = G_N(\mu^*(t_{i-1}))$ and the corrected noise distribution of $(\kappa(s_{i-1})|\hat{X}(t_{i-1}))$. The LNA transition equation (S1 Eq. (4.2)) is then used to derive the distribution of $(\xi(t_i)|\hat{X}(t_{i-1}))$ and the LNA ansatz Eq (3) to obtain $(X(t_i)|\hat{X}(t_{i-1}))$. The measurement equation Eq (12) is finally used to obtain the $(i + 1)$ th term of the likelihood function $P(\hat{X}(t_i)|\hat{X}(t_{i-1}))$ before proceeding to the next iteration. All the distributions obtained in this way are MVN with easily computable parameter values.

If the observations are recorded in short time intervals, the phase correction can be omitted in some steps, in which case the algorithm proceeds as in [35]. Computational methods such as those described in [32] and [35] can then be used to perform likelihood-based statistical inference.

Methods

All computations have been carried out using MATLAB Release 2016b, The MathWorks, Inc., Natick, MA, USA. In particular, the empirical CDF plots, (q-q) plots, histograms, smooth probability density functions and KS distances are derived using the `ecdf`, `qqplot`, `histogram`, `ksdensity`, `kstest` functions of MATLAB and Statistics Toolbox. The computations for the SSA, tau-leap, integration of diffusion and pcLNA simulation algorithms, and the computation of Fisher Information and principal control coefficients for the sensitivity analysis were performed using the PeTTSy software which is discussed in S1 Sect. 14 and is freely available at <http://www2.warwick.ac.uk/fac/sci/systemsbiology/research/software/>. Further details concerning methods are given in S1 Sect. 16.

Discussion

We present a comprehensive treatment of stochastic modelling for large stochastic oscillatory systems. Practical algorithms for fast long-term simulation and likelihood-based statistical inference are provided along with the essential tools for a more analytical study of such systems.

There is considerable scope for future work in various directions. We expect that these results can be extended to a broader class of systems including those that are chaotic in the $\Omega \rightarrow \infty$ limit. Our approach should provide the opportunity to develop new methodology for parameter estimation, likelihood-based inference and experimental design in such systems. Finally, there is currently much interest in information transfer and decision-making in signaling systems and our methods provide new tools with which to tackle problems in this area.

If system biologists are to reliably use complex stochastic models to provide robust understanding it is crucial that there are analytical tools to enable a rigorous assessment of the quality and selection of these models and their fit to current biological knowledge and data. Our aim in this paper is to contribute to that but the results should be of much broader interest.

Supporting information

S1 Appendix. Technical details. In this note we give further details about the mathematical underpinnings of the pcLNA methods discussed in the main paper. (PDF)

S2 Appendix. *Drosophila* circadian clock system. In this note we give details about the *Drosophila* circadian clock and use this system to illustrate further the accuracy of distributions and simulations discussed in the main paper.
(PDF)

S3 Appendix. Brusselator and NF- κ B systems. In this note we give details about the deterministic and stochastic models of the Brusselator and NF- κ B systems and use them to illustrate further the results described in the main paper.
(PDF)

Author Contributions

Conceptualization: Giorgos Minas, David A. Rand.

Data curation: Giorgos Minas.

Formal analysis: Giorgos Minas, David A. Rand.

Funding acquisition: David A. Rand.

Investigation: David A. Rand.

Methodology: Giorgos Minas, David A. Rand.

Project administration: David A. Rand.

Resources: Giorgos Minas, David A. Rand.

Software: Giorgos Minas, David A. Rand.

Supervision: David A. Rand.

Validation: Giorgos Minas, David A. Rand.

Visualization: Giorgos Minas, David A. Rand.

Writing – original draft: Giorgos Minas, David A. Rand.

Writing – review & editing: Giorgos Minas, David A. Rand.

References

1. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 1977 Dec; 81(25):2340–2361. <https://doi.org/10.1021/j100540a008>
2. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. The Journal of Chemical Physics. 2001 Jul; 115(4):1716–1733. <https://doi.org/10.1063/1.1378322>
3. Gillespie DT, Petzold LR. Improved leap-size selection for accelerated stochastic simulation. The Journal of Chemical Physics. 2003 Oct; 119(16):8229–8234. <https://doi.org/10.1063/1.1613254>
4. Gillespie DT. The chemical Langevin equation. The Journal of Chemical Physics. 2000 Jul; 113(1): 297–306. <https://doi.org/10.1063/1.481811>
5. van Kampen NG. Stochastic Processes in Physics and Chemistry, Third Edition. Amsterdam: Elsevier. Boston and London: Elsevier; 2007.
6. Kurtz TG. Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes. Journal of Applied Probability. 1971 Jun; 8(2), 344–356. <https://doi.org/10.1017/S002190020003535X>
7. Kurtz TG. Approximation of Population Processes. Regional Conference Series in Applied Mathematics vol. 36. Society for Industrial and Applied Mathematics. 1981.
8. Hayot F, Jayaprakash C. The linear noise approximation for molecular fluctuations within cells. Physical Biology. 2004 Dec; 1(3–4):205–10. <https://doi.org/10.1088/1478-3967/1/4/002> PMID: 16204840
9. Boland RP, Galla T, McKane AJ. How limit cycles and quasi-cycles are related in systems with intrinsic noise. Journal of Statistical Mechanics: Theory and Experiment. 2008 Sep;P09001.

10. Koepl H, Hafner M, Ganguly A, Mehrotra A. Deterministic characterization of phase noise in biomolecular oscillators. *Physical Biology*. 2011 Oct; 8(5):055008. <https://doi.org/10.1088/1478-3975/8/5/055008> PMID: 21832803
11. Tomita K, Ohta T, Tomita H. Irreversible Circulation and Orbital Revolution: Hard Mode Instability in Far-from-Equilibrium Situation. *Progress of Theoretical Physics*. 1974 Dec; 52(6):1744–1765. <https://doi.org/10.1143/PTP.52.1744>
12. Scott M, Ingalls B, Kærn M (2006). Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. *Chaos*. 2006 Jun; 16(2):026107. <https://doi.org/10.1063/1.2211787> PMID: 16822039
13. Ito Y, Uchida K. Formulas for intrinsic noise evaluation in oscillatory genetic networks. *Journal of Theoretical Biology*. 2010 Aug; 267(2):223–234. <https://doi.org/10.1016/j.jtbi.2010.08.025> PMID: 20800602
14. Schwabedal JTC, Pikovsky A. Phase Description of Stochastic Oscillations. *Physical Review Letters*. 2013 May; 110(20):204102. <https://doi.org/10.1103/PhysRevLett.110.204102> PMID: 25167416
15. Thomas PJ and Lindner B. Asymptotic Phase for Stochastic Oscillators. *Physical Review Letters*. 2014 Dec; 113(25):254101. <https://doi.org/10.1103/PhysRevLett.113.254101> PMID: 25554883
16. Gonze D, Halloy J, Leloup JC, Goldbeter A. Stochastic model for circadian rhythms: effect of molecular noise on periodic and chaotic behaviour. *Comptes Rendus Biologies*. 2003 Apr; 326(2):189–203. [https://doi.org/10.1016/S1631-0691\(03\)00016-7](https://doi.org/10.1016/S1631-0691(03)00016-7) PMID: 12754937
17. Ashall L, Horton CA, Nelson DE, Paszek P, Harper CV, Sillitoe K, Ryan S, Spiller DG, Unitt JF, Broomhead DS, Kell DB, Rand DA, Sée V, White MRH. Pulsatile stimulation determines timing and specificity of NF-kappa B-dependent transcription. *Science*. 2009 Apr; 324 (5924):242–6. <https://doi.org/10.1126/science.1164860> PMID: 19359585
18. Leloup JC, Goldbeter A. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J Biol Rhythms* 1998 Feb; 13(1):70–87. <https://doi.org/10.1177/074873098128999934> PMID: 9486845
19. Hartman P. Ordinary differential equations. New York: John Wiley & Sons, Inc. London and Sydney: Wiley; 1964.
20. Guckenheimer J, Holmes PJ. Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector-fields. New York: Springer-Verlag, Inc. Berlin, Heidelberg and Tokyo: Springer-Verlag; 1983
21. Gupta A, Hepp B, Khammash M. Noise Induces the Population-Level Entrainment of Incoherent, Uncoupled Intracellular Oscillators. *Cell Systems*. 2016 3(6):521–531.e13. <https://doi.org/10.1016/j.cels.2016.10.006> PMID: 27818082
22. Komorowski M, Costa MJ, Rand DA Stumpf MPH. Sensitivity, robustness and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences of the United States of America*. 2011 May; 108(21):8645–8650. <https://doi.org/10.1073/pnas.1015814108> PMID: 21551095
23. Anderson DF. An Efficient Finite Difference Method for Parameter Sensitivities of Continuous Time Markov Chains. *ArXiv e-prints*. 2011 Sep.
24. Srivastava R, Anderson DF, Rawlings JB. Comparison of finite difference based methods to obtain sensitivities of stochastic chemical kinetic models. *The Journal of Chemical Physics*. 2013; 138(7):074110. <https://doi.org/10.1063/1.4790650> PMID: 23445000
25. Brown KS, Sethna JP. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E*. 2003 Aug; 68(2 Pt 1):021904. <https://doi.org/10.1103/PhysRevE.68.021904>
26. Brown KS, Hill CC, Calero GA, Myers CR, Lee KH, Sethna JP, Cerione RA. The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Physical biology*. 2003 Dec; 1(3):184–195. <https://doi.org/10.1088/1478-3967/1/3/006>
27. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*. 2007 Oct; 3(10):1871–78. <https://doi.org/10.1371/journal.pcbi.0030189> PMID: 17922568
28. Rand DA, Shulgin BV, Salazar D, Millar AJ. Design principles underlying circadian clocks. *Journal of The Royal Society Interface*. 2004 Nov; 1(1):119–30. <https://doi.org/10.1098/rsif.2004.0014>
29. Rand DA, Shulgin BV, Salazar JD, Millar AJ. Uncovering the design principles of circadian clocks: Mathematical analysis of flexibility and evolutionary goals. *Journal of Theoretical Biology*. 2006 Feb; 238(3):616–35. <https://doi.org/10.1016/j.jtbi.2005.06.026> PMID: 16111710
30. Rand DA. Mapping the global sensitivity of cellular network dynamics: Sensitivity heat maps and a global summation law. *Journal of The Royal Society Interface*. 2008 Aug; 5 Suppl 1:S59–69. <https://doi.org/10.1098/rsif.2008.0084.focus>
31. Waterfall JJ, Casey FP, Gutenkunst RN, Brown KS, Myers CR, Brouwer PW, Elser V, Sethna JP. Sloppy-Model Universality Class and the Vandermonde Matrix. *Phys. Rev. Lett*. 2006 Oct; 97(15):150601. <https://doi.org/10.1103/PhysRevLett.97.150601> PMID: 17155311

32. Wilkinson DJ. Stochastic modelling for systems biology. Boca Raton: CRC Press/Taylor & Francis; 2012.
33. Kloeden PE, Platen E, Schurz H. Numerical solution of SDE through computer experiments/ Diskette. Berlin: Springer-Verlag; 1997.
34. Iacus SM. Simulation and Inference for Stochastic Differential Equations: With R Examples. New York, NY: Springer New York: Springer e-books; 2008
35. Finkenstädt B, Woodcock DJ, Komorowski M, Harper CV, Davis JRE, White MRH, Rand DA. Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. The Annals of Applied Statistics. 2013 Dec; 7(4):1960–1982. <https://doi.org/10.1214/13-AOAS669>